The NIEHS/NTP-Perlegen Resequencing Project

Project Leader
Frank M. Johnson, Ph.D.

Introduction
In 2004, the NIEHS/NTP embarked on a project to determine the genomic DNA sequence of 15 inbred strains of mice. The project is being conducted under contract to Perlegen Sciences in Mountain View, California. The NTP is contemplating how to make best use of the findings to the benefit of toxicology, the testing of chemicals, and public health.

The results of this project will help identify genes in mice that underlie susceptibility to adverse health conditions such as cancer and heart disease. The results can also help identify genetic factors responsible for variability in the response to toxic agents and help explain why some genotypes may be more susceptible than others to the harmful effects of exposure. Virtually all toxicological testing conducted today fails to take genetic variability in the toxic response into proper account.

Personnel
Key project personnel were Drs. Kelly Frazer, David Cox, Erica Beilharz at Perlegen Sciences, Dr. Molly Bogue at the Jackson Laboratory, Dr. Mark Daly at the Broad Institute, and Dr. Eleazar Eskin at UCLA. Dr. Frank Johnson wrote the scope of work and managed the scientific/technical aspects of the contract as the Government Project Officer.

Objectives and Accomplishments
Project tasks were organized into two phases. Under phase 1 the objectives were:

1. Resequence the genomes of 15 inbred mouse stains using the C57BL/6J as a template.
2. Organize the sequence by chromosome and chromosome location.
3. Identify the variations in sequence that distinguishes the strains.
4. Develop a website to make the data available to the scientific community and the public.
5. Submit the data to the national nucleotide data repositories NCBI and dbSNP.
6. Make the methodology publically available through the website, including the long-range PCR primers and PCR conditions.
7. Analyze the haplotype structure of the strains and identify shared segments.
8. Prepare progress reports and publish the results.

These tasks were completed and published in *Nature*, Frazer *et al*., 2007.

At the same time, an independent group used this data, which was freely available through the NTP-Perlegen website, to describe the subspecific origin of the laboratory mouse published in *Nature Genetics*, Yang *et al*., 2007. That study also resequenced specific genomic segments and improved upon the estimated false positive/ false

negative SNP call rates in our data. In early October 2009, a search of the ISI Web of Science showed our *Nature* 2007 paper to have been cited 88 times.

For Phase 2 the tasks were:

1. Impute the genotypes of 8.27 million SNPs in 40 additional mouse strains using SNPs identified at the Broad Institute and at Perlegen in separate studies.
2. Expand the imputation to include ~100 strains using other data as may be available or become available.
3. Make the imputation data available to the public.
4. Assist/facilitate integration of the SNP genotype data into the JAX phenome database.
5. Develop a research publication describing the results.

This work is described in a manuscript accepted for publication in *Genetics*, Kirby *et al.,* 2009.

The Mouse Strains

The 15 strains selected for the project include 11 commonly used inbred laboratory strains and 4 inbred wild derived strains representing the main *Mus musculuus* subspecies, CAST/EiJ (*Mus musculus castaneus)*, MOLF/EiJ (*Mus musculus molossimus*, PWD/PhJ (*Mus musculus musculus)*, and *WSB/EiJ (Mus musculus domestics)*. Of all subspecies, *domesticus* is by far the most widely distributed geographically. All commonly inbred laboratory strains are genetic mosaics of the various subspecies with origins centuries ago in fancier's breeding stocks, collectors, and pet stores. Although the available inbreds represent a tremendous range of genetic variability, there remains a question as to how representative this variation is with respect to the variability present in natural populations where survival depends on coping with difficult conditions, including toxic exposures. Thus, for resistance genes, new strains derived from the wild may be a better source than the available inbreds.

Sequencing Technology

To conduct the sequencing, 25-mer oligonucleotide probes were synthesized as arrays or features on glass wafers. These arrays were created photolithographically from the known published genomic sequence of the C57BL/6J reference strain (Mouse Genome Sequencing Consortium, *Nature* 2002). 241,806 long-range PCR (LR-PCR) primer pairs were used to amplify the genomes of each of the 15 strains. The amplicons ranged in size from 3 kb to 12 kb, with an average of 10,336 bp. (The PCR failure rate was approximately 5%). The amplicons were then fragmented with DNAse 1 to a peak fragment size of 100 bp and end-labeled with either biotin or fluorescein. The probes were then used to interrogate the unknown sequence represented in the fragments of DNA obtained from the amplified regions of the 15 mouse strains. To perform the interrogation, the amplified DNA fragments were hybridized to the arrays. Then, after washing and staining for the detection of the biotin- and fluorescein-labeled hybridized targets, the arrays were scanned using custom-built confocal scanners and the fluorescence intensity data captured for analysis.

The analysis effectively reaches only unique or nonrepetitive transcribed DNA sequences roughly amounting to 70% or so of the genome. A base-calling algorithm was used that minimized false positives at the expense of permitting substantial false negative calls. As a result, there are undoubtedly many more SNPs existing among the strains than have been identified.

As remarkable as this technology appeared to be when this project was conceived in 2002, it is today obsolete. The cost of our resequencing project was approximately $16 million or ~$1 million per genome, and originally it was estimated to provide ~80% coverage and 1.5X redundancy. By comparison, in June 2009, Illumina announced the availability of a personal genome sequencing service offering “complete” genome analysis of an individual person’s DNA for $48,000 at 30X coverage. Also, in September 2009, Complete Genomics announced plans to offer human genome sequencing at a cost as low as $5,000 per 40X genome.

Recently scientists at NC State University and Baylor College of Medicine proposed the sequencing of a *D. melanogaster* genetic reference panel of 192 wild-derived lines (from natural Raleigh populations) that have been inbred to homozygosity. Extensive information on complex trait phenotypes is being collected on these lines. The project will create a community resource for association mapping of quantitative trait loci that determine various human health related conditions including longevity, body size, and social behavior. Already dozens of associates and collaborators appear to be involved in this project. The organizers estimate the cost of sequencing 192 inbred strains of Drosophila at 10 – 12X to be around $30,000 per strain, using both the Illumina/ Solexa and 454 Sequencing technololgies (Mackay, Richards and Gibbs (2009). This cost is roughly the same as that for individual human genome analysis, and prospects are that costs will continue to decline. A $5000 individual mouse genome may be in reach before too long. Until then, a powerful high-density genotyping array has recently became available (based in part on the SNPs we identified) that contains 623,124 SNPs and 916,269 invariant probes (Yang et al 2009). The array will be useful for characterizing individuals from inbred strains, crosses, and wild populations.

Similar to the proposed project to sequence multiple strains of *Drosophila*, the NTP mouse sequencing project was also established as a community resource, and it has served that purpose well with numerous researchers around the world citing our data in their publications.

Future Plans

For the past 30 years, the NTP has invested resources in investigating alternative model systems, especially alternatives for the rodent cancer bioassay, but there has been little progress (Bucher and Portier, 2004; Collins et al., 2008). Generally, the rodent bioassay uses one strain of mouse and one strain of rat to test an agent, but the strains change over time. For example, the characteristics of both the Fischer 344 rat and B6C3F1 mouse strains have changed, making it necessary for the NTP to consider changing to other strains (Haseman *et al.,* 1998, King-Herbert and Thayer, 2006).

It is desirable to create a public resource consisting initially of 50-100 newly derived wild inbred mouse strains. The available inbred strains may not adequately represent genetic variation existing in nature. Patterns of genomic variability and genotype-environment relationships may occur in natural populations of mice, especially *M. m. domesticus*, the subspecies with the greatest geographic distribution. Investigation of natural mouse populations exposed to various pesticides and other environmental toxins would aid identification of alleles resistant to environmental exposures.

It would also be useful to hold a conference at NIEHS in 2010 to discuss and refine the goals outlined. We would develop a conference agenda and obtain representation from the mouse genetics/genomics community as well as the regulatory science community. The product of this conference would be a consensus document stating the goals, approaches, and relevance of a quantitative genetic/ genomic approach to toxicology.

References

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Bucher, JR and Portier C. (2004) Human Carcinogenic Risk Evaluation, Part V: The National Toxicology Program Vision for Assessing the Human Carcinogenic Hazard of Chemicals. Toxicological Sciences 82: 363–366.

Collins, FS, Gray GM and Bucher JR. (2008) Toxicology: Transforming Environmental Health Protection. Science 319: 906-907.

Haseman, JK, Hailey JR and Morris RW. (1998) Spontaneous neoplasm incidences in Fischer 344 rats and B6C3F1 mice in two-year carcinogenicity studies: A National Toxicology Program update. Toxicologic Pathology 26: 428-441.

Kelly A. Frazer, Eleazar Eskin, Hyun Min Kang, Molly A. Bogue, David A. Hinds, Erica J. Beilharz, Robert V. Gupta, Julie Montgomery, Matt M. Morenzoni, Geoffrey B. Nilsen, Charit L. Pethiyagoda, Laura L. Stuve, Frank M. Johnson, Mark J. Daly, Claire M. Wade and David R. Cox. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448(7157):1050-1053..

King-Herbert A, Thayer K. (2006)**.** NTP workshop: animal models for the NTP rodent cancer bioassay: stocks and strains--should we switch? Toxicol Pathol 34:802-805.

Kirby, A, Kang HM, Wade CM, Cotsapas CJ, Kostem E, Han B, Rivas M, Bogue MA, Frazer KA, Johnson FM, Beliharz EJ, Cox DR, Eskin E and M Daly. (2009) A high density haplotype resource of 94 inbred mouse strains. Genetics (accepted).

Mackay, T, Richards, S and Gibbs R. 2009. Proposal to sequence a Drosophila genetic reference panel: A community resource for the study of genotypic and phenotypic variation.

http://flybase.org/static_pages/news/whitepapers/Drosophila_Genetic_Reference_Panel_Whitepaper.pdf

Yang, H Bell TA, Churchill GA and Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nature Genetics 39:1100-1107. 2007.

Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, Pardo-Manuel de Villena F and Churchill GA. 2009. A customized and versatile high-density genotyping array for the mouse. Nature Methods 6:663-666.